Audiovisual Speech Perception with Degraded Auditory Cues

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation with distinction in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

by

Elizabeth Anderson

The Ohio State University
June 2006

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

Speech perception, although generally assumed to be a primarily auditory process, also depends on visual cues. Audio and visual signals are not only used together when signals are compromised, such as in a noisy environment, but also when the signals are completely intelligible. McGurk and MacDonald (1976) demonstrated the integration of these cues in a paradigm known today as the McGurk effect.

One possible underlying explanation for the McGurk effect is the substantial redundancy in the auditory speech signal. An unanswered question concerns the circumstances that promote optimal perception of auditory and visual signals; is integration improved when one or both signals contain some ambiguity, or is a certain degree of redundancy necessary for integration to occur? If so, how much redundancy is necessary for optimal integration?

The present study began to examine the amount of redundancy necessary for optimal auditory + visual integration. Audio portions of speech recordings were degraded using a software program that reduced the speech signals to four spectral bands, effectively reducing the redundancy of the auditory signal. Performance of participants under four conditions; 1) degraded auditory only, 2) visual only, 3) degraded auditory + visual, and 4) non-degraded auditory + visual, was explored to assess the degree of integration when the redundancy of the auditory signal is reduced. Integration was determined by; 1) comparing the percent of integration across degraded and non-degraded auditory + visual conditions to degraded-auditory only and visual only conditions, and 2) recording the percent of degraded auditory + visual McGurk responses. Results indicate that reducing the redundancy of the auditory signal has no significant effect on auditory + visual integration, suggesting that the amount of redundancy in the auditory signal does not influence the degree of multimodal integration.

Acknowledgments

I would like to thank my advisor, Dr. Janet M. Weisenberger, for providing me with the opportunity to work alongside her on this thesis. I was able to grow personally and professionally through her support, insights, and experience. I am also very grateful to Natalie Feleppelle for her assistance with all aspects of this project. Furthermore, I would like to thank my family and friends for their constant care and encouragement.

This project was supported by an ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship.

Table of Contents

Abstract…….....2

Acknowledgments……..3

Table of Contents……...4

Chapter 1: Introduction and Literature Review……..5

Chapter 2: Method……..15

Chapter 3: Results and Discussion…….22

Chapter 4: Summary and Conclusion…….……26

Chapter 5: References…….28

List of Figures…….30

Figures 1 – 5……...…….……...31

Chapter 1: Introduction and Literature Review

Speech perception, although generally assumed to be a primarily auditory process, also depends on visual cues. In situations where auditory cues are compromised, such as in noisy environments or in hearing-impaired individuals, speech perception can be greatly impaired. Visual input in situations where auditory cues are distorted can significantly improve speech intelligibility. However, there is additional evidence that visual input plays a role even when auditory input is perfect. McGurk and MacDonald (1976) conducted a study to demonstrate the integration of these cues, known today as the McGurk Effect. For this study, auditory syllables were dubbed onto a videotape of a woman's vocalization of contrasting syllables. Subjects were asked to repeat what they heard in both auditory-visual and auditory only conditions. Results of this study indicate that when an auditory [ba] is dubbed onto a visual [ga], participants reported perceiving [da], a fusion of the places of articulation of the two sounds. This result indicates that even when the auditory input is completely intelligible, listeners use information from the visual signal to identify the syllable (McGurk and MacDonald, 1976).

Auditory Cues for Speech Perception

Although speech is an inherently visual process, auditory cues provide beneficial information for perceiving speech sounds. An auditory sound provides information relating to the place, manner, and voicing of a speech sound. This information comes from both spectral and temporal (envelope) aspects of the speech waveform. It has been argued that the speech waveform is highly "redundant" – that is, that the waveform contains far more information than is minimally necessary to identify the presented sound. Even a relatively small amount of temporal and spectral information is useful in identifying speech sounds. Evidence from one

study conducted by Shannon *et al.* (1998) suggests that when the spectral information is reduced to two broad noise bands, which are then modulated by the envelope characteristics of the original speech waveform, the ability to recognize vowels and consonants is impressive. With the use of four bands of noise, speech recognition improves dramatically, providing observers with the ability to recognize almost all of the information provided for manner and voicing. Further, Shannon assessed the effect of spectral warping of envelope cues to test previous research by Drullman *et al.* (1994), which indicated that the reduction of temporal and spectral cues in conditions where speech cues are significantly distorted affects both consonants and vowels (Shannon *et al*., 1998). Shannon *et al.'s* experiment confirmed that consonant recognition is not affected as much as vowel recognition when spectral cues are distorted (Shannon *et al.,* 1998). These studies provide evidence of the robustness and redundancy of the speech signal.

According to McGurk and McDonald, in the absence of auditory cues, visual cues provided by lip movements can be misinterpreted. Nonetheless, they are an important source of speech information (McGurk and McDonald, 1976). In the next section some of the visual cues for speech recognition are described.

Visual Cues for Speech Perception

Previous research suggests that in normal auditory and auditory-visual signals, significant information regarding articulatory features such as place, manner, and voicing is presented to convey speech. Visual signals, however, only provide information regarding place of articulation. In a visual-only situation, individuals must rely on cues given by the talker's visible cues. These cues provide significant information regarding speech and can be obtained from the

movement of the talker's eyes, mouth, and head (Munhall *et al.,* 2004). However, when speech sounds have similar visual characteristics, lack of an auditory cue makes distinguishing the sound a more difficult task for the individual. Speech sounds that have similar visual movement patterns are divided into groups known as visemes. Although viseme groups are beneficial cues for speechreaders, these groups only allow speechreaders to distinguish between groups of sounds, rather than distinction of individual sounds within the group (Jackson, 1988). For example, the viseme group, /p, b, m/, consists of bilabial stops, which are all produced by similar movements of the lips moving together, thus making it difficult to make a distinction between each of the sounds in the group.

Viseme groups are determined by several factors beyond the visual attributes of speech sounds. Differences in articulation patterns among talkers and the environment in which the sounds are produced are substantial elements that contribute to visual speech perception. Talker differences appear to account for significant variations in viseme categories. As discussed by Jackson, talkers that are easy to speechread give rise to more viseme categories than talkers who are more difficult to understand. Additionally, viseme groups that are said to be "universal" are prominent among easy-to-speechread talkers. In contrast, hard-to-speechread talkers provide a smaller number of viseme groups (Jackson, 1988).

In addition, Nitchie (as cited in Jackson, 1988) provided the term *homophenous* to describe speech sounds that appeared alike, but noted that visual cues alone could not provide speechreaders with the necessary information to make a distinction. Homophenous groups consist of sounds that have the same place of articulation, but vary in their voicing and/or nasality. S*peechreading movement*, the classification system described by Jeffers and Barley (as cited in Jackson, 1988), categorizes speech sounds into groups of sounds that have similar visual

patterns. Sounds within a speechreading movement are similar in their motor pattern, but are not visually identical. Unlike consonants, no two vowels share the same visual characteristics; therefore, each vowel has its own distinct articulation pattern (Jackson, 1988). Although these categories provide significant visual information for the speechreader, visual characteristics alone do not provide enough information to determine viseme groupings.

In a study on consonant confusion in consonant-vowel syllables in visual-only, auditory-only, and auditory-visual conditions, Binnie, Montgomery, and Jackson (as cited in Jackson, 1988), indicated that in visual-only conditions the strongest feature for speech perception was place of articulation. Further evaluation of classification systems provided researchers with evidence that /p,b,m/, /f,v/, and /ɵ/, are commonly grouped as visemes, most likely due to visible movements that are *universal* (Jackson, 1988).

Similar to consonants, the production of vowels can provide visual cues that are beneficial for identifying speech sounds. Although every vowel has a distinct shape, vowels can also be classified into visemes. Common viseme groups that provide speechreaders with cues for speech perception of vowels that include extended-rounded contrast and vertical lip separation. While visual components are beneficial for speechreaders, differences among talkers can create confusion among viseme categories for vowels.

Auditory-Visual Integration Theories

Several models of auditory-visual speech integration have been developed to determine the ability to integrate modalities for optimal speech perception. One researcher, Braida (as cited in Grant, 2002), produced the pre-Labeling Model of Integration (PRE), which is used to predict auditory-visual recognition. All information obtained from both auditory and visual conditions

is sustained, meaning neither of the conditions experience interference or biasing from the other condition. Additionally, recognition of auditory-visual cues should be equivalent to or greater weight than auditory or visual cues alone. When scores for auditory-visual integration are similar to the prediction derived from the model then the individual is integrating efficiently, rehabilitation should be focused on increasing scores for auditory or visual recognition. In contrast, individuals performing below predicted auditory-visual scores are not integrating efficiently and should receive integration training in order to improve their scores. Further analysis of the PRE model has shown that individual hearing-impaired subjects have been significantly over-predicted; thus, these individuals should benefit considerably from rehabilitation for integration of auditory and visual cues (Grant, 2002).

Another theory for explaining auditory-visual integration is the Fuzzy Logical Model of Perception (FLMP). Massaro (as cited in Grant, 2002) constructed this model of integration efficiency to reduce the variation between the predicted auditory-visual and the obtained auditory-alone and visual-alone speech recognition scores. According to Grant, the FLMP underestimates the integration ability of humans; thus, the discrepancy creates a doubt as to whether the FLMP is a reliable measurement of integration efficiency (Grant, 2002).

Furthermore, Grant and Seitz (as cited in Grant, 2000) found that by watching the lip and face movements, detection of speech can be improved up to 3dB. From these results, it was implied that subjects are able to correlate the obtained visual and acoustic information. With a correlation greater than .9 and an amplitude envelope at its maximum, a positive effect on speech detection thresholds will occur; this effect is known as the Bimodal Coherence Masking Protection (BCMP). According to Grant and Seitz (as cited in Grant, 2000), this effect indicates that visual information will partially guard target speech signals from the effects of noisy

environments. Additional examination of this effect was done by Grant to determine whether or not speechreading can support auditory detection of verbal sentences when the peaks of the amplitude envelope correspond with a temporal location correlated to the area of lip opening and the amplitude envelope. Subjects were introduced to two sentences that were digitally bandpass-filtered and centered in masking noise; one sentence was centered on the first formant (F1) speech region (100-800 Hz) and the other sentence was centered on the second formant (F2) speech region (800-2200 Hz). The subjects were instructed to identify the period that contained the target sentence at varying levels of intensity. Results of this test implied that BCMP magnitude is dependent on the speech signal's temporal and spectral characteristics. Moreover, the study provides further support that speechreading can provide information to cue listeners about when and where to expect a signal based on the movements of the speaker's lips, thus improving the listener's ability to identify speech in noisy environments (Grant, 2000).

Role of Redundancy in Audiovisual Speech Perception

Through the integration of auditory cues and visual cues, individuals can be less dependent on auditory cues when the cue is distorted by noise or a reduced acoustic signal; visual cues under these conditions can significantly increase the intelligibility of the signal (Munhall *et al.*, 2004). Further evidence suggests that auditory speech signals are highly redundant. In a study conducted by Shannon *et al.*, speech recognition of consonants, vowels, and sentences was measured when the spectral distribution of envelope cues was distorted. The results of this study supported previous studies in that when spectral cues were distorted, consonant recognition was consistently less sensitive to the distortions than vowel recognition. Consequently, poor vowel recognition led to a total disruption in sentence recognition. The

finding provided by Shannon *et al*. indicates that phonemes can be perceived with high accuracy even when spectral cues are absent from a stimulus.

Furthermore, even when speech is reduced to three sinusoids, speech sounds can be discriminated and perceived to provide essential information for lip readers. Additional information provided by this research is helpful in understanding the ability for high levels of speech recognition in cochlear implant patients. Although only a few electrodes are used to stimulate neurons with a portion of a speech signal, evidence suggests that even when only four electrodes are used, cochlear implant patients are able to perceive speech. Additionally, for speech materials processed through four bandpass filters, it was concluded that although the alignment of the frequency of the analysis bands and carrier bands is crucial in providing good performance, reducing the redundancy in the acoustic signal does not impair the ability for an individual to identify speech (Shannon *et al.*, 1998). These results suggest that the tonotopic distribution is imperative for speech recognition in conditions where the envelope cues are distorted (Shannon *et al.,* 1998).

Similarly, in a previous study by Shannon *et al.* (1995), the ability to recognize speech with reduced spectral information was examined. Temporal envelopes of speech were manipulated to preserve temporal cues of the spectral band while reducing its spectral information. Results of this study showed that although the spectral content was greatly reduced, speech recognition increased as the number of noise bands increased and with only three bands of modulated noise, a high level of speech recognition was still achieved (Shannon *et al.,* 1995). In addition, this experiment indicates that a surplus of information for a speech sound is supplied by non-distorted auditory speech signals.

Impoverished auditory signals can become highly intelligible when presented with visual cues in connected speech. In an attempt to determine the significance of differing integration abilities among individuals in predicting auditory-visual consonant and sentence recognition among individuals, Ken Grant and Philip Seitz (1998) compared several auditory-visual integration measures. Congruent and discrepant auditory-visual nonsense syllable and sentence recognition tasks were employed in the integration measures compared in the study. Natural speech that consists of a single sound source that is in synchrony with the visual signal is said to be congruent. In contrast, the discrepant materials are created by dubbing an auditory production of one sound onto a visual production of a different sound, as demonstrated by McGurk and McDonald. The results of the research conducted by Grant and Seitz showed that even a highly impoverished auditory speech input (i.e. Fθ, or fundamental frequency) led to intelligible sentences when visual cues were added (Grant and Seitz, 1998). However, the lack of correlation between the benefits provided in connected sentences and in isolated nonsense syllables suggests that participants were employing "top-down" cognitive processing, such as knowledge of the language, in the connected speech situation. The conditions under which benefit from added visual input is obtained for isolated syllables require further investigation.

One of the most impoverished, and hence least redundant, speech signals is sine-wave speech. Remez (1981) and his colleagues generated a three time-varying sinusoid that reflected a naturally produced utterance; all of the acoustic cues of traditional speech were absent from the stimuli. According to Remez *et al.*, the stimulus sentence should be perceived as three separate tones. Three conditions were tested for this study; individual groups of subject listeners received varying levels of information regarding the stimuli being presented. The first two groups were given very little information regarding the stimuli and the third group was told exactly what the

sentence was that they would hear. The results of this study indicated that even naïve listeners can detect the linguistic content of an utterance in time-varying sinusoids without traditional acoustic cues (Remez *et al.,* 1981).

Determining how the degree of redundancy in auditory speech signals affects the strength of the McGurk effect is important in understanding the role of redundancy in auditory-visual speech perception. Even though both auditory and visual signals provide information regarding the place of articulation of the speech stimulus, the auditory speech signal is highly redundant, whereas the visual speech signal can be rather ambiguous. The possibility exists that by stripping the redundancy from the auditory speech signal, the strength of the McGurk effect could be reduced; thus, forcing individuals to rely on visual cues for optimal perception. One unanswered question concerns the circumstances that promote optimal perception of auditory and visual signals; is integration improved when one or both signals contain some ambiguity, or is a certain degree of redundancy necessary for integration to occur? If so, how much redundancy is necessary for optimal integration? To investigate this question it is important to determine how the integration process is affected when a vast amount of the redundancy from the auditory signal is reduced. Understanding of the role of redundancy of the auditory speech signal in the auditory-visual speech process has important implications for intervention for individuals with auditory impairments.

Thus, evidence in the literature suggests that audiovisual integration can be highly beneficial in speech identification when the auditory signal is degraded in some manner. In addition, in their non-degraded form speech signals are highly redundant. Finally, the McGurk effect demonstrates that individuals use both visual and auditory cues in deciphering speech even

when the auditory signal is not degraded. However, the existence of the McGurk effect may depend on a certain degree of redundancy in the auditory signal.

The present study investigated the question of how audiovisual integration occurs for isolated syllables by presenting highly reduced, non-redundant auditory speech cues, specifically degraded speech, together with visual speech information. A group of normal hearing adult subjects were asked to identify speech stimuli under conditions where both auditory and visual components represent the same speech sound, as well as conditions where auditory and visual components represent two different speech sounds. Results of this study should have implications for signal processing strategies for hearing aids and cochlear implants as well as for designs for rehabilitation programs.

Chapter 2: Method

Participants

Ten adult female college students, ages 21-23, participated in this study. All participants reported normal hearing and normal vision. Five of the ten participants had completed undergraduate courses in phonetics, while the other five participants had not taken courses containing information on phonetics and language. Participants received $30.00 for their involvement in this study.

Interfaces for Stimulus Presentation

Visual Presentation

Presentation of degraded auditory and visual stimuli was similar for all participants. Each participant was tested with stimuli under four conditions: 1) visual only; 2) degraded auditory only; 3) visual plus degraded auditory; and 4) visual plus normal auditory. Under each condition, participants sat in a chair inside the chamber, with the window shade pulled up for visual access to the video monitor and with the door to the chamber sealed shut. A 50 cm video monitor was placed about 60 cm outside the window of a sound attenuated chamber. The monitor was positioned at eye level, about 4 feet away from the participant's head. Each participant was presented with stimuli consisting of several talkers under several conditions, all of which were randomized. Stimuli were presented using recorded DVDs on the video monitor for each condition. For the visual only condition, the video monitor's sound was turned off.

Degraded Auditory Presentation

The degraded auditory stimuli were presented from the headphone output of the video monitor to Sennheiser, 600-ohm circum-aural headphones. Under the degraded auditory only condition, the shade above the chamber window was pulled down, in order to remove visual cues

from the video monitor's screen. When visual and degraded auditory stimuli were presented together, the shade above the window was raised up to allow viewing of the video monitor screen.

Stimuli

Stimulus Selection

A set of eight CVC syllables were used as the stimulus for this study. Each syllable was selected in accordance with the following conditions:

1.) Pairs of the stimuli were minimal pairs, differing by only one phoneme, the initial consonant

2.) All stimuli were accompanied by the vowel /ae/, since it does not involve lip rounding or lip extension

3.) Multiple stimuli were used in each category of articulation, including: place (bilabial, alveolar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced)

4.) All stimuli were presented without a carrier phrase (citation style)

5.) Stimuli were known to elicit McGurk-like responses

Stimuli

For each condition, the same sets of stimuli were administered. The 8 stimuli used were as follows:

1.) mat

2.) bat

3.) gat

4.) pat

5.) cat

6.) zat

7.) sat

8.) tat

Stimulus Presentation

Audio Signal Degrading

Seven talkers provided the speech stimuli for the auditory stimuli. Each talker was recorded through a microphone directly into a computer, using the software program Video Explosion Deluxe, which permitted files to be stored in .wav format. Each talker repeated a selected set of eight monosyllabic stimuli words, five times each. These auditory files were then input to a subroutine created by Bertrand Delgutte in MATLAB 5.3, a computer software program. The subroutine ("chimeras") begins with two stimuli, one the input speech waveform, and the other a broadband noise. The program swaps the waveform and fine structure of the two stimuli. Each speech signal was then filtered into four broad spectral bands. The bandwidths of the four channels are chosen to provide equal spacing in basilar membrane distance. The upper cutoff frequencies for the four spectral bands were: 504 Hz, 1,794 Hz, 5,716 Hz, and 17,640 Hz. For the present study, the waveform containing the noise fine structure and the temporal envelope cues of the original speech waveform was preserved for use, and the other waveform was discarded. The resulting auditory stimulus was thus similar to those created by Shannon *et al.* (1998), as described in the Introduction.

Digital Video Editing

Visual stimuli for the study were obtained by first recording seven male and female talkers with a digital video camera; each talker repeated the list of 8 stimulus words 5 times. Stimuli from the recordings were then downloaded and edited using a computer software program, Video Explosion Deluxe. Within this program, auditory stimuli created with the

"chimeras" subroutine can be dubbed onto the visual representation of a speech sound. Thus, it was possible to create audio-visual stimuli that featured both normal auditory and degraded auditory components. It was also possible to create stimuli that feature a different auditory and visual syllable (dual-syllable stimuli), thus permitting the analysis of McGurk-type integration effects. For the present study, the visual stimuli produced by a talker were paired only with auditory stimuli produced by that same talker.

Through the use of another computer software program, Sonic MY DVD, stimulus lists were created and burned onto recordable DVDs. Multiple DVDs were produced for each talker for each condition, all with different randomized stimulus orders, in order to minimize the possibility of effects that can occur from order of stimulus presentation. For this study, the DVDs were presented on a DVD player connected to the video monitor.

The testing was broken into four presentation conditions: visual only, degraded auditory only, visual plus degraded auditory, and visual plus normal auditory; each participant was tested under all four conditions, with the order of conditions randomized across participants. For each trial, participants were asked to repeat the word that they thought had been presented. These responses were then recorded by an experimenter conducting the testing.

Visual Alone: Under the visual alone portion of this study, participants were presented with visual stimuli from the recorded DVDs, played and visible from the video monitor. They were all asked to say the word they felt was being said by the talker. Because this condition required that all auditory cues were absent, the participant did not wear headphones and the video monitor's sound was turned off. For this condition, participants were seated inside the sound attenuating chamber, facing the video monitor placed outside the chamber window.

Degraded Auditory Alone: Under the auditory alone condition, participants wore headphones in the sound attenuating chamber, which allowed the degraded auditory stimuli from the video monitor to be heard. Randomized orders of DVDs were played for each of the participants. Participants were also seated in a chair in the back of the chamber, facing the video monitor outside the window. Participants were asked to repeat the word they perceived. For this condition, the shade above the window was pulled down, and the video monitor was turned off in order to remove visual cues of the talker.

Visual plus Degraded Auditory: Under the visual plus degraded auditory condition, participants were again seated in a sound attenuating chamber, facing the video monitor outside the chamber window. Participants wore a set of headphones in order to hear the degraded auditory stimuli, and the shade to the window was pulled up to allow the participant to view the video monitor, which presented the visual stimuli. Again, a random order of DVDs was played for each participant via the video monitor and headphones.

Visual plus Normal Auditory: Testing under the visual plus normal auditory condition was administered in order to compare the performance under normal conditions with the performance with a degraded auditory signal. Participants wore headphones for reception of the degraded auditory stimuli, and again, the shade of the window was pulled up to allow the participant to view the video monitor, which provided the visual stimuli. Under this condition, all participants watched and listened only to talkers 2, 5, and 7 via the video monitor and headphones.

Procedure

Testing Setup

Testing for this study took place in a basement lab room of The Ohio State University's Speech and Hearing Department. The lab provided a quiet environment that was well-lit with its fluorescent lighting. Participants were seated in a chair along the back wall of one of the lab room's single walled, sound-attenuating chambers. All participants sat the same distance away from the chamber's window and the video monitor. Examiner feedback and subject responses were transmitted through an intercom system in the chamber.

A 50 cm video monitor was placed outside the booth approximately 4 feet away from the participant and facing a double-glass window on one wall of the chamber. The video monitor was positioned at eye level for optimal view of the visual stimuli. For all conditions, participants were seated in a chair against the back wall of the chamber, facing the window. The chamber door was completely sealed for all testing, and the window shade was pulled down under the degraded auditory alone condition, and raised for conditions where visual cues were allowed. In the visual alone condition, the video monitor's sound was turned completely off. For conditions where degraded auditory stimuli were presented, participants wore headphones.

Testing Tasks

Under each of the three conditions, participants were presented with 60 randomly-ordered stimulus syllables, which were conveyed by several talkers on prerecorded DVDs. Each stimulus word was presented multiple times, while each list was completed for only one condition, for one participant. The words presented to the participants consisted of eight stimuli, differing only in the initial consonant. However, in portions of the visual alone condition and portions of the degraded auditory conditions, stimuli were used that would provide the opportunity to elicit McGurk-like responses. After each presentation, the participants were asked to provide the examiner with a verbal response of the word they thought they heard based on the

auditory cues presented in the degraded and normal auditory conditions and what they thought was said based on the visual cues presented during the visual only condition. To record the participant responses the examiner used data sheets. The presentation order of each condition was varied across participants.

Testing Presentation: Testing consisted of four conditions, which were all tested using 60 stimuli presented via prerecorded DVDs, consisting of several talkers and randomized. In single modality testing, which included visual alone and degraded auditory alone conditions, the video and audio portions of the stimulus consisted of the same syllable. During the visual plus degraded auditory condition, the 60 stimuli words consisted of 30 "same" trials and 30 "different" trials. All modalities were presented with the same speech stimulus for the "same" trials. In the "different" trials, the visual modality was presented with a different speech stimulus from the auditory modality. Trials were randomized to eliminate the chance of the participants knowing which type of trial was being presented. Risk of memorization by the participants was minimized by producing a substantial set of different DVDs, each consisting of a random stimulus order.

Testing Procedure: Each participant was tested with four stimulus conditions. Under each condition, sixty trials were presented via prerecorded DVDs. Visual plus degraded auditory testing was executed in order to elicit McGurk responses. Every participant was tested under the conditions listed below in randomized order.

Visual only

Degraded auditory only

Visual plus degraded auditory

Visual plus normal auditory

Chapter 3: Results and Discussion

Results were analyzed for two types of stimuli. First, performance was assessed for single-syllable presentations, in which all modalities tested (degraded auditory only, visual only, degraded auditory + visual, normal auditory + visual) received the same stimulus. For these stimuli, percent correct identification performance was measured. Integration can be assessed by determining the degree to which degraded auditory + visual performance was better than performance in the degraded auditory only or visual only conditions.

Second, performance was evaluated for dual-syllable presentation, in which each modality received a different stimulus (e.g., auditory + visual testing, with an auditory stimulus of "bat" and a visual stimulus of "gat."). For these stimuli, there is no single "correct" response, so responses are categorized as "visual," "auditory," or "other." Integration for these stimuli is defined as a response that is different from either the visual or the auditory stimulus.

Percent Correct Identification

Figure 1 shows the percent correct identification for degraded auditory stimuli in auditory only, visual only, and auditory + visual conditions. The figure indicates that degrading the speech signal into four channels did decrease intelligibility. Furthermore, the results show that listeners were able to integrate the visual and degraded auditory signals to achieve higher performance in the auditory + visual condition.

For comparison purposes, three of the seven talkers (2,5, & 7) also produced stimuli under normal auditory conditions. Figure 2 shows a comparison of the percent correct identification scores for three talkers between the normal and degraded auditory signal in the auditory + visual condition. All three talkers produced near perfect listener identification under normal auditory conditions, and dropped to about 75-80% correct identification in the degraded

auditory condition. Perfect performance in the normal auditory condition is not surprising because the auditory signal and visual signal were uncompromised, and therefore contained all the necessary information for identification.

For the remainder of the analysis performance was averaged across talkers, as well as across observers. Figure 3 shows the percent correct identification for normal versus degraded auditory conditions in auditory + visual testing. As previously indicated, performance was near perfect in normal auditory conditions, but significantly lower in the degraded auditory condition. Statistical analysis using a dependent groups t-test indicated that the difference was significant [$t(9) = -7.05$, $p < .05$]. This result indicates that the four-channel broad spectral degrading performed on these stimuli does indeed remove information from the auditory signal.

McGurk Type Integration

For the remainder of the analysis only trials in which a different auditory and visual stimulus was presented were included. In these trials there is no "correct" response. Observer responses were broken down to reflect the percent of time observers chose the visual stimulus, the percent of time observers chose the auditory stimulus, and the percent of time observers chose an "other" response, which reflects integration of the visual and auditory stimulus. These responses are shown in Figure 4 for normal and degraded auditory conditions. Not surprisingly, in the normal auditory condition observers showed a heavy reliance on auditory responses and provided relatively few visual responses. As previously mentioned in the Introduction, the auditory stimulus is highly redundant, whereas the visual stimulus can be ambiguous because only the place of articulation can be determined from it. Interestingly, this pattern is completely reversed in the degraded auditory condition, where reducing the redundancy in the auditory signal led to a sharp decrease in auditory responses and a sharp increase in visual responses.

Thus, despite the relative ambiguity of the visual stimulus, observers were more likely to rely on it for information, compared to the degraded auditory stimulus. This appears to be true even though performance in the auditory only condition was higher than that in the visual only condition when percent correct performance was measured, indicating that overall more information was available from the degraded auditory stimulus than from the visual stimulus. Thus, it is surprising that observers nonetheless chose the visual stimulus more frequently than the auditory stimulus when the auditory input was degraded.

Overall, however, the percentage of responses that reflect integration is very similar for normal and degraded auditory conditions, suggesting that the degree of integration is not influenced by the amount of redundancy in the auditory signal. Again, statistical analysis showed a significant difference between normal and degraded auditory conditions in the percent of visual responses [$t(9) = 9.48$, $p < .05$] and the percent of auditory responses [$t(9) = -9.08$, $p < .05$]. However, the overall integration was not significantly different in the percent of "other" responses [$t(9) = .77$] .

Although the overall level of McGurk type integration was not affected, the specific types of integration were very different for the two presentation conditions, suggesting that reducing redundancy in speech signals does affect the integration process. Figure 5 shows the classification of "other" responses from Figure 4 to reflect the type of integration observed. Responses were classified under three categories: the first and most common type of integration is a fusion, e.g., when the auditory stimulus presented is /ba/ and the visual stimulus is /ga/, observers often combine the places of articulation and respond with /da/; the second, a combination, e.g., when the auditory stimulus /ga/ is paired with the visual stimulus /ba/, observers often combine the stimuli and respond with /bga/, or; neither fusion or combination,

which occur when an observer is certain about the response, but it is not a true McGurk response (e.g., responses of "hat" or "brat"). As can be seen in Figure 5, observers produced a larger percentage of fusion responses in the normal auditory condition as compared to the degraded auditory condition. Similarly, a greater percentage of combination responses was produced by observers in the normal auditory condition as compared to the degraded auditory condition. However, it should be noted that in either condition, combination responses were minimal. Observers produced a larger amount of responses that were "neither" in the degraded auditory condition. Once again, dependent groups t-tests showed significant differences in fusion responses [$t(9) = -3.62$, $p < .05$]; combination responses [$t(9) = -2.71$, $p < .05$]; and neither responses [$t(9) = 6.58$, $p < .05$] in the type of integration response for the two auditory conditions.

Overall, the degrading method effectively reduced redundancy in speech signals and also appeared to have an effect on listeners' reliance on different modality inputs. While the overall level of McGurk type integration was not affected, the specific types of integration were very different for the two presentation conditions, suggesting that reducing redundancy in speech signals does affect the integration process.

Chapter 4: Summary and Conclusion

Results of this study indicate that broad spectral degrading effectively reduces information available in the speech signal. This is supported by the lower percent correct in the degraded auditory condition, where observers were correct about 60% of the time when presented with single-syllable stimuli. Even so, substantial integration was observed through comparison of the auditory + visual condition (81% correct) with the auditory only condition (60% correct), thus indicating that individuals are able to achieve higher performance by integrating visual signals with degraded auditory signals.

In addition, degrading the auditory signal has important effects on the integration process, as seen with dual-syllable stimuli, where the auditory and visual signals present different syllables. Observers rely on the visual stimuli when the auditory signal is degraded, even though that might not be the optimal strategy, as evidenced by better degraded audio (60%) than visual only (35%) percent correct.

Furthermore, the overall amount of McGurk style integration was similar for normal and degraded auditory conditions, suggesting that observers are trying to use all of the information available to them and integrate, regardless of whether the information is good or less good. The specific type of integration (fusion, combination) also appears to differ when a degraded auditory signal is used. When the auditory signal is normal, observers tend to respond significantly more often with fusion responses and combination responses than when the auditory signal is degraded. However, the percent of "neither" responses is significantly higher when the auditory signal is degraded, than when the signal is normal. In the "neither" responses observers were attempting to integrate the visual and auditory signals, but due to the reduction in the auditory stimulus the responses that were produced were not averages of the places of articulation as

would generally be seen in standard McGurk responses. These results not only indicate that the observers were very flexible in their response strategy and were able to adapt quickly to the change in the quality of auditory input, but also, the reduction of the information in the auditory stimulus produced very different patterns of response. Thus, by reducing the redundancy in the auditory signal, the process by which observers integrate is affected.

Results from this study are just a preliminary look into this issue. The present study examined the effects of reducing the redundancy using four-channel filtered speech, which is only one of several degrading strategies. Therefore, future work should employ different numbers of channels in spectral degrading (e.g. 2-channel, or 8-channel degrading). Additionally, other ways to degrade the signal should be looked at to see if they differentially affect the nature or type of integration. It is imperative that further analysis into this issue be performed before categorically concluding the results indicated by this study.

Finally, results of this study have long-term implications for signal processing strategies for hearing aids and cochlear implants. As indicated by Shannon *et al*. (1998), it seems that even a limited number of channels of speech input is sufficient for some degree of identification. However, the present study suggests that the type of integration of auditory and visual inputs might be different when the auditory input is degraded in some way. Furthermore, the finding that observers rely on visual input even when there might be better information in the auditory input may provide insight into the design of aural rehabilitation programs. It is possible that patients could be taught how to determine the optimal information channel in particular situations, and thus be taught to rely more extensively on this channel.

Chapter 5: References

Grant, K.W. (2000). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America, 109 (5),* 2272 – 2275.

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America, 112 (1),* 30-33.

Grant, K.W. & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104 (4)*, 2438-2449.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90 (5)*, 99-114.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

Munhall, K.G., Kroos, C., Jozan, C., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perceptions & Psychophysics, 66 (4)*, 574 – 583.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech

perception without traditional speech cues. *Science*, *212 (4497*),
947-949.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995).
Speech recognition with primarily temporal cues. *Science*, *270*, 303-304.

Shannon, R.V., Zeng, F.G., Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *The Journal of the Acoustical Society of America, 104 (4),*
2467-2475.

List of Figures

Figure 1: Percent Correct Identification in Impoverished Auditory Conditions

Figure 2: Percent Correct Identification by Talker (Auditory + Visual Testing)

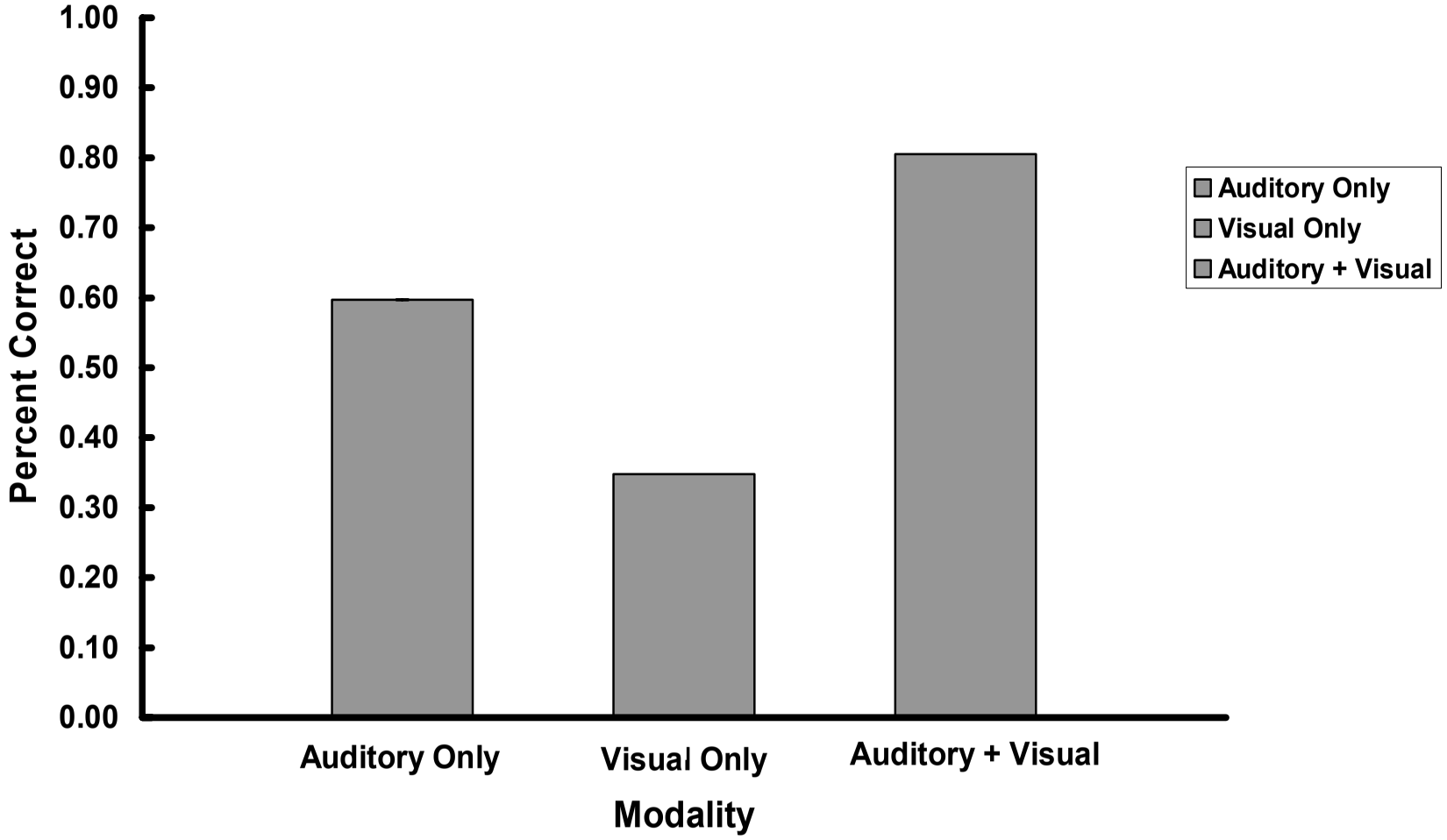
Figure 3: Percent Correct Identification in Normal vs. Impoverished Auditory Conditions (Auditory + Visual Testing)

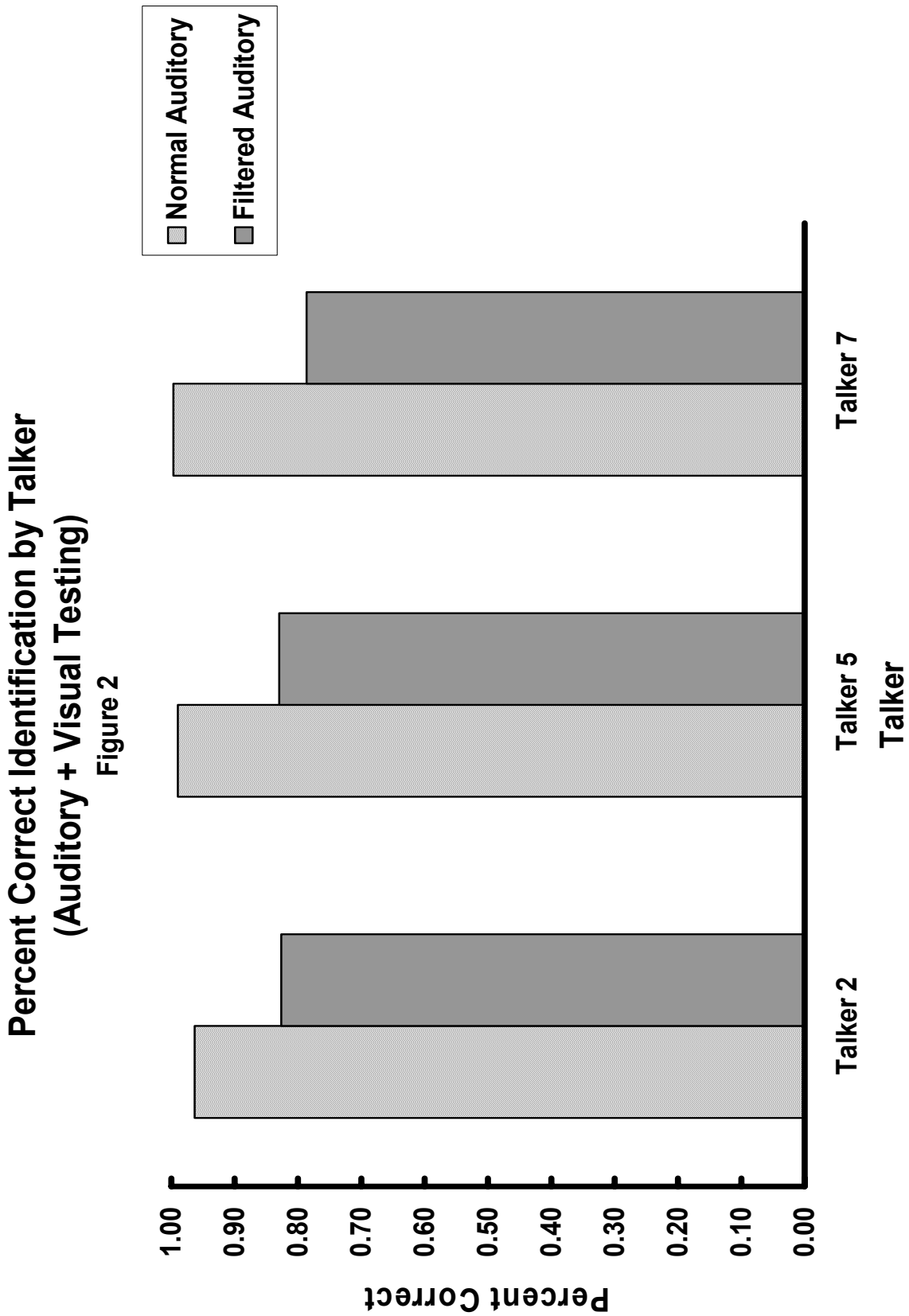
Figure 4: Percent Response in Normal vs. Impoverished Auditory Conditions (Auditory + Visual Testing)

Figure 5: Classification of "Other" Responses in Normal vs. Impoverished Auditory Conditions (Auditory + Visual Testing)

Percent Correct Identification in Impoverished Auditory Conditions

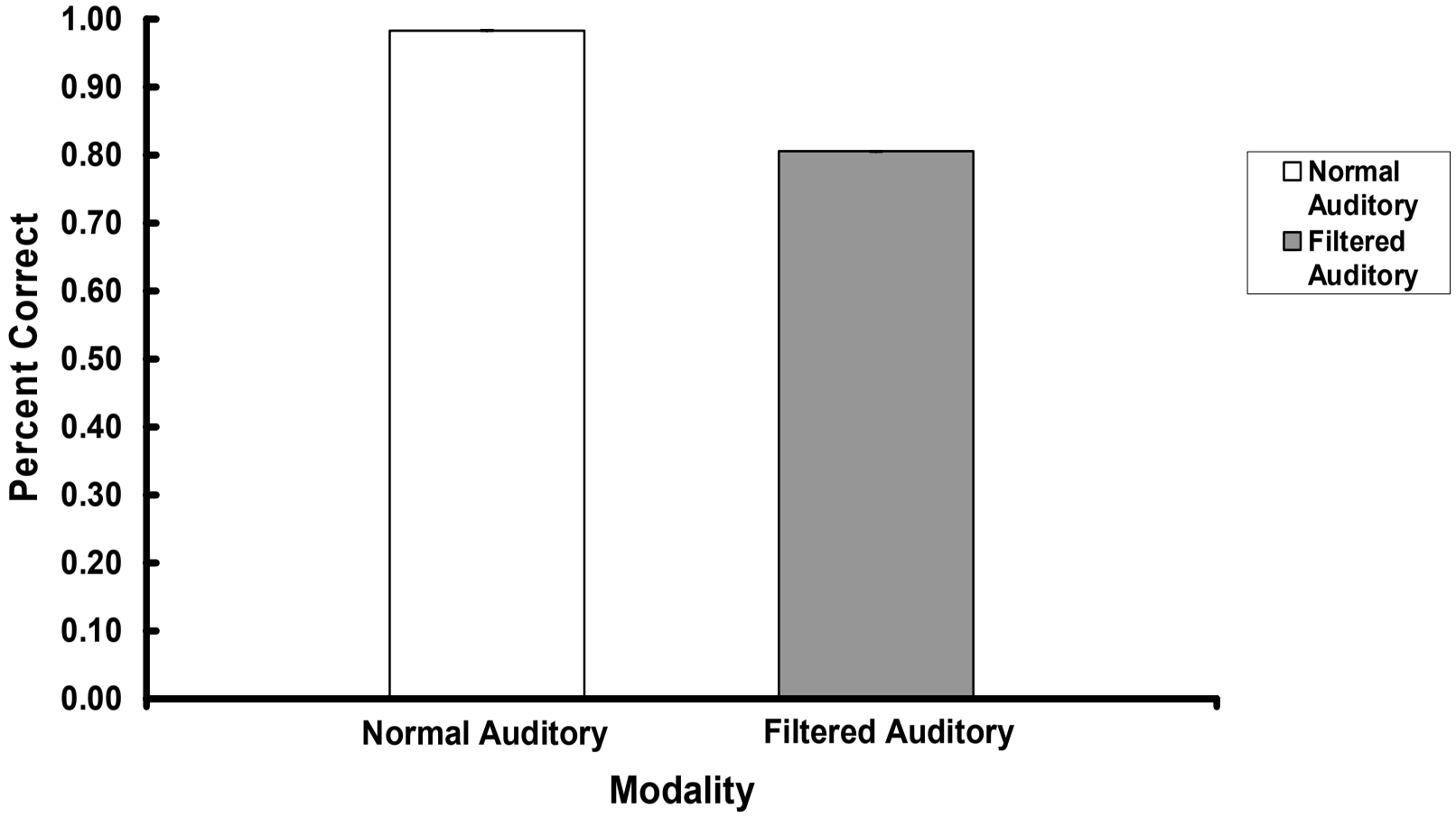
Figure 1




Percent Correct Identification by Talker
(Auditory + Visual Testing)
Figure 2
Normal Auditory
Filtered Auditory
Percent Correct
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Talker 2
Talker 5
Talker 7
Talker

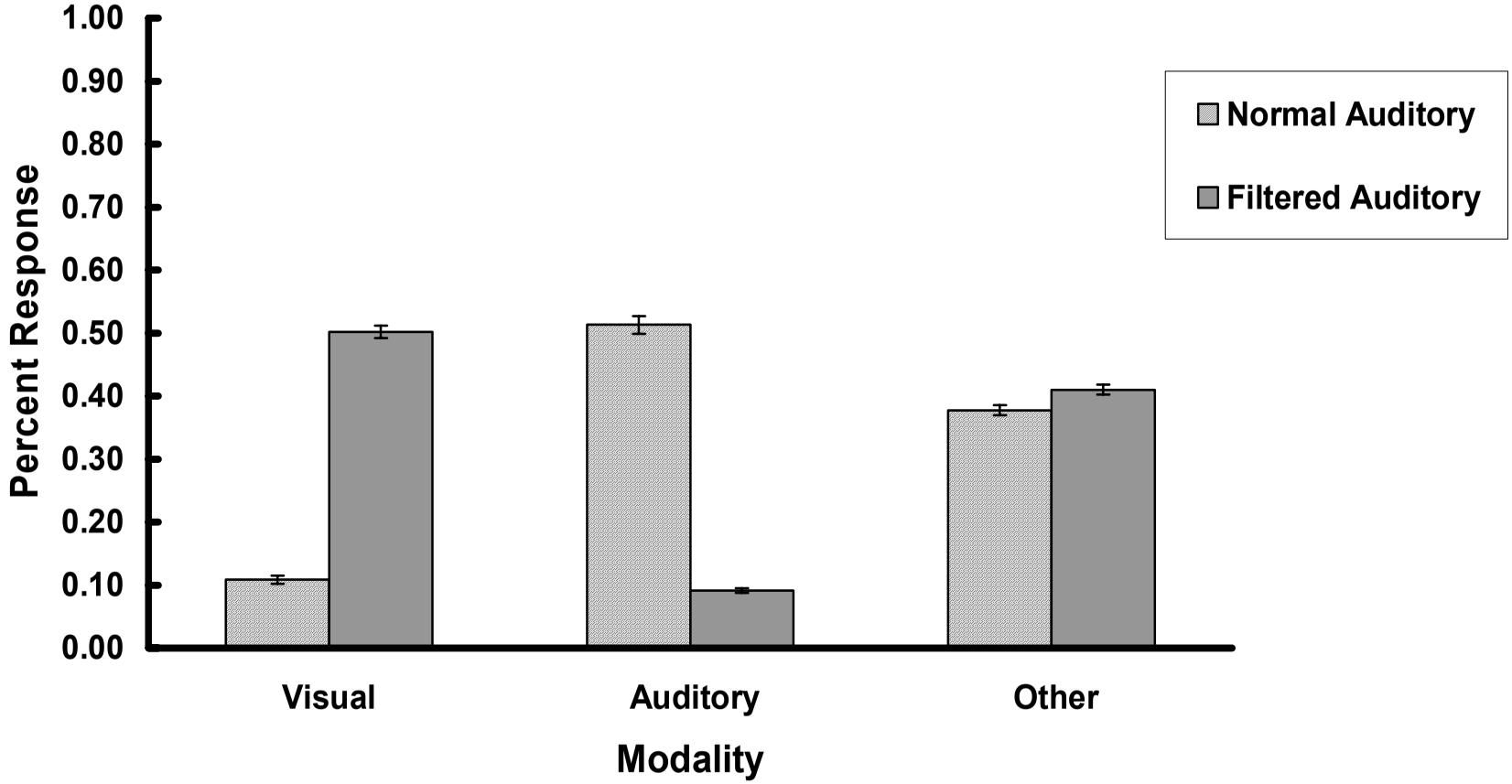
Percent Correct Identification in Normal vs. Impoverished Auditory Conditions (Auditory + Visual Testing)

Figure 3



Percent Response in Normal vs. Impoverished Auditory Integration (Auditory + Visual Testing)

Figure 4



Classification of "Other" Responses in Normal vs. Impoverished Auditory Conditions (Auditory + Visual Testing)

Figure 5